

The logo for iris AperTO, featuring the text "iris AperTO" in white on a red rectangular background.

UNIVERSITÀ
DEGLI STUDI
DI TORINO

This is the author's final version of the contribution published as:

M. Oksnes; S. Bensing; A.L. Hulting; O. Kämpe; A. Hackemann; G. Meyer; K. Badenhoop; C. Betterle; A. Parolo; R. Giordano; A. Falorni; L. Papierska; W. Jeske; A.A. Kasperlik-Zaluska; V.K.Chatterjee; E.S. Husebye; K. Løvås K.. QUALITY-OF-LIFE IN EUROPEAN PATIENTS WITH ADDISON'S DISEASE; VALIDITY OF THE DISEASE-SPECIFIC QUESTIONNAIRE. THE JOURNAL OF CLINICAL ENDOCRINOLOGY AND METABOLISM. 97 pp: 568-576.
DOI: 10.1210/jc.2011-1901

The publisher's version is available at:
<http://press.endocrine.org/doi/abs/10.1210/jc.2011-1901>

When citing, please refer to the published version.

Link to this full text:
<http://hdl.handle.net/2318/92983>

This full text was downloaded from iris - AperTO: <https://iris.unito.it/>

iris - AperTO

University of Turin's Institutional Research Information System and Open Access Institutional Repository

Validity of the disease-specific questionnaire AddiQoL in European patients with Addison's disease

Marianne Øksnes, Sophie Bensing, Anna Lena Hulting, Olle Kämpe, Annika Hackemann, Gesine Meyer, Klaus Badenhoop, Corrado Betterle, Anna Parolo, Roberta Giordano, Alberto Falorni, Lucyna Papierska, Wojciech Jeske, Anna A. Kasperlik-Zaluska, V. Krishna K. Chattterjee, Eystein S. Husebye, Kristian Løvås

Institute of Medicine (M.Ø., K.L., E.S.H.), University of Bergen, Bergen, Norway; Department of Medicine (K.L., E.S.H.), Haukeland University Hospital, Bergen, Norway; Department of Molecular Medicine and Surgery (A.L.H., S.B.), Karolinska Institutet, Stockholm, Sweden; Department of Medical Sciences (S.B., O.K), Uppsala University, Uppsala, Sweden; Division Endocrinology & Diabetology (K.B., A.H., G.M.), Medical Department 1, University Hospital of the Goethe-University Frankfurt, Frankfurt am Main, Germany; Endocrine Unit Department of Medical and Surgical Sciences (C.B., A.P.), University of Padua, Padua, Italy; Department of Clinical and Biological Sciences (R.G.), University of Turin, Turin, Italy; Department of Internal Medicine (A.F.), University of Perugia, 06126 Perugia, Italy; Endocrinology Clinic(A.K.Z., L.P., R.K.), The Medical Centre of Postgraduate Education, Ceglowska 80, Warsaw, Poland; Department of Medicine(V.K.K.C.), University of Cambridge, Cambridge CB22QQ, United Kingdom .

Corresponding author: Marianne Øksnes, Institute of Medicine, Haukeland University Hospital, Jonas Liesvei 65, N-5021 Bergen, Norway. E-mail: marianne.oksnes@med.uib.no. Tel +4755973075, Fax +4755972950

Précis: The Addison's disease-specific quality-of-life questionnaire AddiQoL was tested in five European countries, yielding a revised 30-item questionnaire with high reliability and validity.

Word count: 3 527

Key words: AddiQoL, Health-related quality-of-life, adrenal insufficiency, Addison's disease

This work was supported by the European Union Seventh Framework Program Grant 201167, Euradrenal, and the National Institute of Health Research Cambridge Biomedical Centre (United Kingdom). The Swedish contribution was also supported by the Swedish Society of Medicine.

Disclosure Summary: No conflict of interest was reported by any of the Authors.

Abstract

Context: Patients with Addison's disease (AD) self-report impairment in specific dimensions on well-being questionnaires. An AD disease-specific quality-of-life questionnaire (AddiQoL) was developed to aid evaluation of patients.

Objective: We aimed to translate and determine construct validity, reliability, and concurrent validity of the AddiQoL questionnaire.

Methods: After translation, the final versions were tested in AD patients from Norway (n=107), Sweden (n=101), Italy (n=165), Germany (n=200) and Poland (n=50). Construct validity was examined by exploratory factor analysis (EFA) and Rasch analysis, aiming at unidimensionality and fit to the Rasch model. Reliability was determined by Cronbach's α and Person Separation Index (PSI). Longitudinal reliability was tested by Differential Item Functioning (DIF) in stable patient subgroups. Concurrent validity was examined in Norwegian (n=101) and Swedish (n=107) patients.

Results: EFA and Rasch analysis identified six items with poor psychometric properties. The 30 remaining items fitted the Rasch model and proved unidimensional; supported by appropriate item and person fit residuals and a non-significant Chi-Square probability. Cronbach's α 0.93 and PSI 0.86 indicates high reliability. Longitudinal reliability was excellent. Correlation with SF-36 and PGWB scores were high. A shorter Fatigue subscale comprising eight items also proved valid and reliable. Testing of AddiQoL-30 in this large patient cohort showed significantly worse scores with increasing age, and in women compared with men, but no difference between patients with isolated AD and those with concomitant diseases.

Conclusion: The validation process concluded with a revised 30-item AddiQoL questionnaire and a Fatigue subscale with good psychometric properties and high reliability.

Primary adrenal insufficiency (Addison's disease, AD) is a rare chronic disease treated with glucocorticoid and mineralocorticoid replacement (1), and additional replacement of the adrenal androgen dehydroepiandrosterone (DHEA) is debated (2-5). Novel treatment strategies such as modified-release hydrocortisone tablets or continuous subcutaneous hydrocortisone infusion are under investigation (6-9). There is no 'gold-standard' for assessment of treatment, but a clinical scoring system has been proposed (10). Patient surveys reproducibly report impairment in particular dimensions of general well being questionnaires (3, 11, 12). Generic questionnaires have been applied to study differences between subgroups of patients with AD (13, 14) and other autoimmune endocrinopathies (15). However, questionnaires containing disease-specific items are likely to be more sensitive to effects that clinicians wish to monitor (16). Recently, The AddiQoL 36-item disease-specific questionnaire was developed as an evaluative disease-specific questionnaire in AD (17), which might facilitate the detection of changes in well-being in future clinical trials and during regular follow-up of the patients.

Validity is the process of demonstrating that an instrument measures what it intends to measure, and that it is useful for this purpose. Construct validity aligns a questionnaire to a theorized underlying trait, and involves testing of correlation between the items. Here, we used Rasch analysis to explore the psychometric properties of AddiQoL. Rasch analysis is a mathematical item response model increasingly used in somatic medicine and endocrinology, such as validation of QoL-AGHDA, AcroQoL and CushingQoL (18-20). The objective is to test how well the observed data fit to the expectations of the mathematic measurement model (21, 22).

Reliability implies the degree to which an instrument is free from random error. The traditional reliability coefficient (Cronbach's α) indicates how well an individual item correlates with the other items in a questionnaire. In Rasch analysis, the Person Separation index (PSI) is equivalent to Cronbach's α , and represents the power of the construct to discriminate between respondents, giving an indication of how precisely patients have been spread out along the continuum (23). Test-retest reliability or repeatability is the correlation between scores obtained by the same persons on two separate occasions given that their clinical condition is stable.

Concurrent validity concerns how the questionnaire performs against some "gold standard" instrument, usually by exploring correlation of questionnaire scores. The AddiQoL is to our knowledge the first disease-specific HRQoL questionnaire in AD, therefore no such "gold standard" exist. In endocrinology, the SF-36 (Short-Form-36) and the PGWB (Psychological General Well-Being Index) have been included in the validation of AGHDA (24, 25), AcroQoL (26) and CushingQoL (20).

Validation of a questionnaire requires responses from a number of subjects, which is difficult to obtain within one country for a rare disease such as AD. Hence, in the current study we translated the original English AddiQoL into Norwegian, Swedish, Italian, German and Polish versions; these were administered to large cohorts of patients with AD in each country for evaluation of construct validity and reliability. Test-retest reliability was tested in patient subgroups in Norway, Italy and Sweden. Concurrent validity was investigated by examining correlation between the AddiQoL scores and simultaneously obtained results from SF-36 and PGWB questionnaires in Norway and Sweden. We also sampled AddiQoL data from a random population sample in Norway. The final questionnaire was ultimately used to assess HRQoL in different subgroups of this large cohort.

Methods

Design and subjects

First, the AddiQoL was translated from English into Norwegian, Swedish, German, Polish and Italian versions, following international recommendations (27). Second, patients with verified AD were recruited from patient registries, or consecutively from outpatient clinics. The patients received an inviting letter containing study information and the AddiQoL; by returning the anonymized questionnaires they were included in the study. Patients in Norway and Sweden in addition received the SF-36 and the PGWB for analysis of concurrent validity. There were no exclusion criteria. Patient characteristics such as age, sex and concurrent autoimmune diseases were retrieved from the registries or through an additional questionnaire. For analysis of longitudinal reliability a subgroup of at least 20 patients from Norway, Sweden and Italy received a second AddiQoL 2-6 weeks after the first. For normative data, a random sample from the Norwegian population received an inviting letter containing study information and the AddiQoL; by returning the anonymized questionnaires they were included. Third, the responses from the patients were analyzed by exploratory factor analysis (EFA) and Rasch analysis for assessment of validity and reliability, and hence for amendment of the questionnaire. The study was approved by the Regional Ethics committees in each country.

Translation

The forward translation was performed by a minimum of three native speakers of the target language, who had good knowledge of English. Translations were performed locally by the study group in each country. Their preliminary versions were discussed locally by a panel of experts, i.e. clinicians in endocrinology, agreeing upon versions to be evaluated further. For quality control, these versions were assessed by two professional translators (Lionbridge Technologies Inc.), native speakers of the target language, who evaluated the conceptual equivalence with the original, clarity and use of a familiar register. Thus, two adjusted versions of each AddiQoL translation were generated, which were re-evaluated by the study groups in each country, who concluded upon a final version.

Questionnaires

The original AddiQoL is a 36-item questionnaire; each item contains six scoring categories. Twenty-five items are negative HRQoL statements that need to be reversed for questionnaire scoring; thus a higher score indicates a higher level of HRQoL. The questionnaire was developed in the English language, and initial statistic analysis performed in 85 patients from the UK (17). The SF-36 is a generic HRQoL questionnaire, widely used and thoroughly validated (28, 29). The SF-36 is translated into many languages, and has been used in previous studies of HRQoL in AD (11, 30). The PGWB is a validated 22-item generic HRQoL questionnaire that has been translated into several languages, intended to measure the subjective feeling of psychological well-being (31).

Construct validity and Rasch analysis

The Rasch model rests on the idea that useful measurement involves examination of only one human attribute at a time (unidimensionality), and that a higher score implies more of the measured concept. The model allows quantitative assessment (additivity of items) from data that are ordinal but not necessary linear, based on logistic transformation of the item responses (22). First, EFA was used to examine dimensionality of AddiQoL. Second, we applied Rasch analysis (RUMM 2020 software (32)), to further identify misfitting items, aiming for a unidimensional construct solution with fit to the Rasch model (33). The analysis gives overall fit statistics and estimates the individual items' and persons' fit to the model.

Overall fit statistics include item-person interaction statistics, calculated as mean item location and mean person location. Both person fit and item fit is transformed by RUMM to approximate a Z-score; this represents a standardized Normal Distribution. Therefore, if the items and persons fit the model perfectly, the mean Fit Residual is expected to be zero with standard deviations around 1. Overall fit statistics also includes χ^2 statistics for item-trait test-of-fit to the model. This tests whether the items work as expected at group level along the range of the scale. A non-significant χ^2 probability implies that the hierarchical ordering of items and persons do not vary across the range of the scale.

Individual item fit analysis identifies misfitting or biased items, thus producing information about how to optimize the item solution and refine the questionnaire. Generally, any item with a Fit Residual greater than +/- 2.5 is a cause of concern; a high positive item Fit Residual indicates that the item does not separate well between high and low person ability, and a high negative item Fit Residual indicates redundancy or local dependency (see below) of the item.

Other causes for misfit to the Rasch model are disordered thresholds and differential item functioning (DIF, item bias). A threshold is the point where the probability of endorsing two neighboring response

alternatives is equal; one threshold exists for each transition between one scoring alternative to the next. To obtain ordered thresholds, each item and response alternative was assessed, and collapsed or rescored when necessary to improve fit. DIF analysis explores item performance and instrument performance across different patient groups. DIF exists if one patient group scores significantly different on an item compared with another patient group, given that the overall HRQoL level of the two groups is the same (21). Here, we performed DIF analysis for patient sex, age, concurrent disease, country and time point (for test-retest reliability analysis).

Each item's difficulty (item location) and each person's ability (person location) are organized in ordered hierarchies (22, 34). By plotting item location and person location on the same scale the targeting of the items for the sample population can be explored. A perfect targeting is indicated if average person location is zero.

To produce a psychometrically meaningful total score; the scale has to be unidimensional (22). This is tested by Principal Component analysis of the Fit Residuals. Fit to the model and an absence of a significant pattern among the Fit Residuals supports the scale being unidimensional, i.e. there is only one concept being measured. Local dependency exists when there is covariance between the response patterns of items, and this is considered a breach of the strict unidimensionality that the Rasch model requires. This can be corrected for by grouping items with covariance together, i.e. treating the item group mathematically as a single combined item (35). If fit to the model and unidimensionality are present, the individual Person location can then be used (directly or transformed to a total score) as a psychometrically valid total score for each patient, achieving an interval measure, as opposed to the ordinal raw scores.

Reliability

PSI is calculated as the ratio of adjusted (true) variance to observed variance and represents the proportion of variance that is not due to error (23, 34). A PSI of 0.85 is generally required if the scale is to be used on the individual level. For longitudinal reliability, a test-retest DIF analysis was performed for patient subgroups with stable clinical condition in Norway, Italy and Sweden, over two to six weeks' intervals.

Concurrent validity and normative data

AddiQoL scores were compared with SF-36 scores and PGWB score in Norway and Sweden. Spearman's rho with two-tailed significance was calculated for the correlation analysis. Mann-Whitney U-test was used to compare Norwegian patients' AddiQoL scores with AddiQoL results from a random Norwegian population sample. Comparison of AddiQoL and Fatigue scores in different

subgroups of patients was performed by multiple linear regression analysis with sex, age, country and comorbidity as independent variables.

Results

Subjects

A total of 615 patients were recruited from Norway (n=107), Italy (n=157), Germany (n=200), Sweden (n=101) and Poland (n=50). The Polish data were omitted from some of the analyses due to low number and lacking data. The original UK data were also included in the pooled data. Patient characteristics are presented in Table 1. Information on patient comorbidities was available for Norway, Italy, Germany and Sweden. Test-retests were available from patients in Norway (n=37), Italy (n=25) and Sweden (n=29). SF-36 and PGWB scorings were available from Norway (n=107) and Sweden (n=101). Of the 2000 Norwegians invited to participate in the study as a normative sample, 76 invitations were returned unopened and 539 persons responded, producing a response rate of 28%. Of the respondents 54% (283) were female and 56% (300) were below the age of 50 (18-39, n=166; 40-49, n=134; 50-59, n=123; >60, n=107).

Translation and quality of questionnaire responses

Overall, the evaluations from the professional translators were favorable; mostly minor errors were noted. Most of the suggestions from the translators were endorsed. The item “I feel lightheaded” proved difficult to translate. In Poland the clinicians did not endorse the suggestions from professional translators, as they believed that their version was more in keeping with everyday doctor-patient communication. Overall, the rate of missing responses was below 1%. There was a tendency towards missing responses on page 2 of the questionnaire, as a few patients did not complete page 2. Items regarding sexuality had the most missing responses, “I am satisfied with my sex-life” (item11) 7%, and “I have lost interest in sex” 4.9%. This was not evenly distributed among countries as 22% of patients in Poland did not report on item 11, whereas the result for Norway was 0.9% (UK 7%, Italy 7.3%, Germany 5.5% and Sweden 6.9%).

Construct validity

Initial overall fit statistics of the 36-item questionnaire showed misfit to the Rasch model. EFA identified four sub-dimensions of AddiQoL, which we denoted Fatigue (8 items), Emotions (8 items), Symptoms (11 items) and Miscellaneous (sleep, sexuality and impact of intercurrent disease, 6 items) (see Supplementary Table 1 for item overview). The three items “nocturia”, “dry skin” and “gaining weight” did not belong to any sub-dimension by EFA, had high item fit residuals and poor discriminating properties in the Rasch analysis. Further Rasch analysis revealed that the Fatigue

domain fitted the Rasch model and achieved unidimensionality in all countries, and no significant DIF was present for sex, age or co-morbidity. The Emotions sub-dimension showed overall good fit to the Rasch model, but there was multidimensionality in the Norwegian data. The item “Emotional stress makes me exhausted” had high residual correlation with the item “I cope well in emotional situations” and displayed DIF in the Swedish data, hence the former item was discarded, improving fit. In the Symptoms sub-dimension the item “I have salt cravings” seems to have high clinical relevance but nevertheless displayed misfit in nearly all countries (Figure 1a). Elimination of this item improved fit to the model, but χ^2 probability remained significant in Italian and German data. The Miscellaneous sub-dimension showed overall good fit and unidimensionality, but the item “I have lost interest in sex” displayed DIF by sex in the Norwegian, Italian and German data, and DIF by age in the Italian and German data (Figure 1b). Removal of this item improved fit.

Disordered thresholds were present for many items, indicating that the subjects had difficulties differentiating between some response alternatives. We found that rescoring the original six response alternatives (123456) to four (122334) by collapsing the scoring categories “a little of the time”/“some of the time”, “a good bit of the time”/“most of the time”, “agree”/“slightly agree” and “disagree”/“slightly disagree” improved fit and produced ordered thresholds.

The 30 remaining items, rearranged in the four revised sub-dimensions as super-items, fitted the Rasch model; supported by a non-significant item-trait interaction ($\chi^2=0.56$) in the pooled data. Also, this item solution proved unidimensional. Table 2 displays the overall fit statistics for the pooled data and for individual countries. There was no significant DIF between genders, and no DIF when comparing results from patients with isolated AD with patients with autoimmune poly-endocrine syndromes. Significant DIF for age was present in the Emotions sub-dimension in the Swedish and the pooled data. Significant DIF for country was present in the Fatigue, the Symptom and the Miscellaneous sub-dimensions. Based on these results we chose to go ahead with validation of the revised 30-item questionnaire (AddiQoL-30) and the Fatigue subscale.

Targeting of the items to the total patient population is shown in Figure 2. Mean person location was 0.21, indicating that mean patient score was slightly higher than the HRQoL level targeted by the mean of the items (set at zero). Mean person location for individual countries was -0.04 (Poland), 0.09 (Germany), 0.14 (UK), 0.21 (Italy), 0.23 (Norway) and 0.27 (Sweden), indicating good targeting in all countries.

Reliability

AddiQoL-30 demonstrated good reliability as indicated by Cronbach’s α 0.93 and PSI 0.86. PSIs for individual countries are presented in Table 2. For the Fatigue sub-dimension PSI ranged from 0.89 to

0.91 in individual countries, indicating excellent reliability as a separate scale. A total of 91 clinically stable patients from Norway, Sweden and Italy performed test-retest 2-6 weeks after the first evaluation. Longitudinal reliability was excellent, as no significant DIF between separate time points was detected.

Concurrent validity, patient scores and normative data

Rasch-transformed AddiQoL-30 and Fatigue scores were compared with SF-36 scores and PGWB scores in Norwegian (n=107) and Swedish (n=101) patients. Results from the correlation analyses are given in Table 3. AddiQoL raw scores in patients (median 89, n= 99) were significantly lower than in controls in Norway (Median 97, n = 462; U = 14799, z = -5.516, p <0.001, r = 0.23), shown for individual age groups and sex in Figure 3. Rasch transformed scores from all countries are shown in Figure 4. Regression analysis adjusting for age, sex, country and comorbidity showed that women scored significantly worse than men (AddiQoL-30 p <0.001; Fatigue p= 0.001) and demonstrated worse scores with increasing age (AddiQoL-30 p<0,001; Fatigue p= 0.001), as indicated for Norwegian data in Figure 4. No statistical difference was found between patients with isolated AD and those with autoimmune polyendocrine syndromes.

Discussion

The validation process concluded with a revised 30-item AddiQoL questionnaire. High reliability, here demonstrated by adequate PSI and a high Cronbach's α , indicates that the items discriminate well between groups of patients with different HRQoL levels. High reliability and good discriminative properties will usually imply good evaluative properties. On the other hand, a too high reliability coefficient is not necessarily desirable, as the same properties that increase the reliability coefficients might reduce ability to detect change (responsiveness) (36).

The final revised AddiQoL-30 fitted the stringent Rasch model, which implies that basic requirements for a measurement instrument such as unidimensionality, order, and additivity are fulfilled.

This solution had the best fit to the Rasch model, proved unidimensional and had the best targeting to the patient sample. The Fatigue sub-dimension alone had good psychometric properties, and a higher reliability than AddiQoL-30, and was therefore kept as a separate index for further validation.

However, the Fatigue scale did not target the whole patient population as well as AddiQoL-30, yielding risk of floor- and ceiling effects that might compromise responsiveness.

One important validity aspect is whether individual items work similarly in different patient subgroups, that is, whether item bias exists (37). No significant DIF was found between patients with isolated AD and patients with autoimmune polyendocrine syndromes, and no DIF by sex remained in AddiQoL-30, indicating equal performance regardless of comorbidities and sex. Thus reassured that

AddiQoL-30 performs equally in these patient groups, the significantly lower total scores in females than males represent a true difference. Similarly, no significant difference in AddiQoL-30 scores between patients with isolated AD and those with polyendocrine syndromes is reliable. Both findings are consistent with earlier studies applying SF-36 in AD (11, 14). We cannot rule out that the demonstrated DIF by country is due to qualitative differences of the translations. If the aim were to study HRQoL differences between countries, statistical adjustment is required for items with DIF country (38). However, DIF country may have negligible clinical impact on the HRQoL results from a clinical trial (39).

We demonstrate high correlation between the AddiQoL-30 score and SF-36 and PGWB. For SF-36 the correlation was highest with the vitality and general health scales, which were also most affected in previous studies in AD (3, 11, 14, 30). Normative data are not essential in the validation of a disease-specific questionnaire, since several of the issues may not be relevant to healthy subjects. Normative data were only collected from Norway; the response rate was low, which could imply selection bias, but age and sex distribution resembled that of the patient group. We found a statistically significant difference between the patients and the controls, but the effect size was small. Several of the items showed ceiling effects in the controls, which underestimates the effect size. Furthermore, comparison of patients with healthy controls always implies some response bias or a response shift due to adaptation to chronic diseases (40).

The results of the Rasch analysis suggest a revised scoring algorithm. The analysis revealed that the six response categories of each item had to be collapsed into four to obtain order, additivity, and fit to the model. The unidimensional structure of the rescored AddiQoL-30 suggests that an index based on the algebraic sum of the item (after reversal of negative items) scores will be valid for practical purposes, for instance in cross-sectional studies. AddiQoL was however primarily developed as an evaluative instrument of within-individual measurement in clinical trials. For this purpose Rasch transformed person location scores will be the optimal psychometric solution. AddiQoL-30 could possibly be further shortened to reduce the respondents' burden. However, the items most sensitive to change might not be the most well-fitting items (36, 41), hence further item reduction will be re-evaluated after testing for responsiveness. The Fatigue sub-dimension displayed optimal overall fit statistics and could be useful as a separate AddiQoL short version.

In conclusion, the validation process concluded with a revised 30-item AddiQoL questionnaire and a Fatigue subscale that have high internal consistency and reliability. Its validity as a HRQoL instrument in AD was further substantiated by high correlation with SF-36 subscales and the PGWB Index. Although further studies are necessary to examine its responsiveness to changes in HRQoL over time,

this study suggests that the AddiQoL could become a valuable tool in the assessment of subjective health status in patients with Addison's disease.

Acknowledgments

In Norway, we thank Inger Johanne Næss for handling the questionnaires and contributing endocrinologist Martina Moter Erichsen. In Sweden, we thank Lena Ehrenstig for distributing the questionnaires. In Italy, we thank Stefania Marzotti in Perugia and Daniela Forno in Turin for assistance in recruiting patients. In Germany, we thank contributing endocrinologists (Reinhard Santen, Alexander Mann, Holger Willenberg, Stefanie Hahner, Bruno Allolio, Nicole Reisch, and Endokrinologikum Hamburg) and the German Endocrine Society's adrenal division. We also thank the patient organization GLANDULA. We also thank all the patients with Addison's disease who participated.

References

1. Arlt W, Allolio B 2003 Adrenal insufficiency. *Lancet* 361:1881–1893
2. Løvås K, Gebre-Medhin G, Trovik TS, Fougner KJ, Uhlving S, Nedrebø BG, Myking OL, Kampe O, Husebye ES 2003 Replacement of dehydroepiandrosterone in adrenal failure: no benefit for subjective health status and sexuality in a 9-month, randomized, parallel group clinical trial. *J Clin Endocrinol Metab* 88:1112–1118
3. Gurnell EM, Hunt PJ, Curran SE, Conway CL, Pullenayegum EM, Huppert FA, Compston JE, Herbert J, Chatterjee VK 2008 Longterm DHEA replacement in primary adrenal insufficiency: a randomized, controlled trial. *J Clin Endocrinol Metab* 93:400–409
4. Arlt W, Callies F, van Vlijmen JC, Koehler I, Reincke M, Bidlingmaier M, Huebner D, Oettel M, Ernst M, Schulte HM, Allolio B 1999 Dehydroepiandrosterone replacement in women with adrenal insufficiency. *N Engl J Med* 341:1013–1020
5. Hunt PJ, Gurnell EM, Huppert FA, Richards C, Prevost AT, Wass JA, Herbert J, Chatterjee VK 2000 Improvement in mood and fatigue after dehydroepiandrosterone replacement in Addison's disease in a randomized, double blind trial. *J Clin Endocrinol Metab* 85:4650–4656
6. Løvås K, Husebye ES 2007 Continuous subcutaneous hydrocortisone infusion in Addison's disease. *Eur J Endocrinol* 157:109–112
7. Johannsson G, Filipsson H, Bergthorsdottir R, Lennernas H, Skrtic S 2007 Long-acting hydrocortisone for glucocorticoid replacement therapy. *Horm Res* 68(Suppl 5):182–188
8. Newell-Price J, Whiteman M, Rostami-Hodjegan A, Darzy K, Shalet S, Tucker GT, Ross RJ 2008 Modified-release hydrocortisone for circadian therapy: a proof-of-principle study in dexamethasone-suppressed normal volunteers. *Clin Endocrinol (Oxf)* 68:130–135
9. Johannsson G, Bergthorsdottir R, Nilsson AG, Lennernas H, Hedner T, Skrtic S 2009 Improving glucocorticoid replacement therapy using a novel modified-release hydrocortisone tablet: a pharmacokinetic study. *Eur J Endocrinol* 161:119–130
10. Arlt W, Rosenthal C, Hahner S, Allolio B 2006 Quality of glucocorticoid replacement in adrenal insufficiency: clinical assessment

- vs. timed serum cortisol measurements. *Clin Endocrinol (Oxf)* 64: 384–389
11. Hahner S, Loeffler M, Fassnacht M, Weismann D, Koschker AC, Quinkler M, Decker O, Arlt W, Allolio B 2007 Impaired subjective health status in 256 patients with adrenal insufficiency on standard therapy based on cross-sectional analysis. *J Clin Endocrinol Metab* 92:3912–3922
 12. Løvas K, Loge JH, Husebye ES 2002 Subjective health status in Norwegian patients with Addison's disease. *Clin Endocrinol (Oxf)* 56:581–588
 13. Bleicken B, Hahner S, Loeffler M, Ventz M, Allolio B, Quinkler M 2008 Impaired subjective health status in chronic adrenal insufficiency: impact of different glucocorticoid replacement regimens. *Eur J Endocrinol* 159:811–817
 14. Erichsen MM, Løvås K, Skinningsrud B, Wolff AB, Undlien DE, Svartberg J, Fougner KJ, Berg TJ, Bollerslev J, Mella B, Carlson JA, Erlich H, Husebye ES 2009 Clinical, immunological, and genetic features of autoimmune primary adrenal insufficiency: observations from a Norwegian registry. *J Clin Endocrinol Metab* 94:4882–4890
 15. Ott J, Promberger R, Kober F, Neuhold N, Tea M, Huber JC, Hermann M 2011 Hashimoto's thyroiditis affects symptom load and quality of life unrelated to hypothyroidism: a prospective case-control study in women undergoing thyroidectomy for benign goiter. *Thyroid* 21:161–167
 16. Prieto L, Santed R, Cobo E, Alonso J 1999 A new measure for assessing the health-related quality of life of patients with vertigo, dizziness or imbalance: the VDI questionnaire. *Qual Life Res* 8:131–139
 17. Løvås K, Curran S, Øksnes M, Husebye ES, Huppert FA, Chatterjee VK 2010 Development of a disease-specific quality of life questionnaire in Addison's disease. *J Clin Endocrinol Metab* 95:545–551
 18. Wiren L, Whalley D, McKenna S, Wilhelmsen L 2000 Application of a disease-specific, quality-of-life measure (QoL-AGHDA) in growth hormone-deficient adults and a random population sample in Sweden: validation of the measure by rasch analysis. *Clin Endocrinol (Oxf)* 52:143–152
 19. Webb SM, Prieto L, Badia X, Albareda M, Catala M, Gaztambide S,

- Lucas T, Paramo C, Pico A, Lucas A, Halperin I, Obiols G, Astorga R 2002 Acromegaly Quality of Life Questionnaire (ACROQOL) a new health-related quality of life questionnaire for patients with acromegaly: development and psychometric properties. *Clin Endocrinol (Oxf)* 57:251–258
20. Webb SM, Badia X, Barahona MJ, Colao A, Strasburger CJ, Tabarin A, van Aken MO, Pivonello R, Stalla G, Lamberts SW, Glusman JE 2008 Evaluation of health-related quality of life in patients with Cushing's syndrome with a new questionnaire. *Eur J Endocrinol* 158:623–630
21. Pallant JF, Tennant A 2007 An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol* 46:1–18
22. Tennant A, McKenna SP, Hagell P 2004 Application of Rasch analysis in the development and application of quality of life instruments. *Value Health* 7(Suppl 1):S22–S26
23. Tennant A, Conaghan PG 2007 The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum* 57:1358–1362
24. McKenna SP, Doward LC, Alonso J, Kohlmann T, Niero M, Prieto L, Wíren L 1999 The QoL-AGHDA: an instrument for the assessment of quality of life in adults with growth hormone deficiency. *Qual Life Res* 8:373–383
25. Suzukamo Y, Noguchi H, Takahashi N, Shimatsu A, Chihara K, Green J, Fukuhara S 2006 Validation of the Japanese version of the Quality of Life-Assessment of Growth Hormone Deficiency in Adults (QoL-AGHDA). *Growth Horm IGF Res* 16:340–347
26. Webb SM 2006 Quality of life in acromegaly. *Neuroendocrinology* 83:224–229
27. Acquadro C, Conway K, Hareendran A, Aaronson N 2008 Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value Health* 11:509–521
28. McHorney CA, Ware Jr JE, Raczek AE 1993 The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs.

Med Care 31:247–263

29. Dupyu H ed 1984 The Psychological General Well-Being (PGWB) Index. New York: Le Jaque Publishing

30. Andrich L, Sheridan, Luo 2003 RUMM2020. Perth, Australia: RUMM Laboratory

31. Smith RM 2000 Fit analysis in latent trait measurement models. J Appl Meas 1:199–218

32. Smith Jr EV 2001 Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. J Appl Meas 2:281–311

33. Tang K, Beaton DE, Lacaille D, Gignac MA, Zhang W, Anis AH, Bombardier C 2010 The Work Instability Scale for Rheumatoid Arthritis (RA-WIS): Does it work in osteoarthritis? Qual Life Res 19:1057–1068

34. Teresi JA 2006 Overview of quantitative measurement methods. J Clin Endocrinol Metab, February 2012, 97(2):568–576 jcem.endojournals.org 575
The Endocrine Society. Downloaded from press.endocrine.org by [\${individualUser.displayName}] on 10 January 2015. at 08:16 For personal use only. No other uses without permission. . All rights reserved.

Equivalence, invariance, and differential item functioning in health applications. Med Care 44:S39–S49

35. Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, Lawton G, Simone A, Carter J, Lundgren-Nilsson A, Tripolski M, Ring H, Biering-Sorensen F, Marincek C, Burger H, Phillips S 2004 Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care 42:I37–I48

36. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, Gundy C, Koller M, Petersen MA, Sprangers MA 2009 The practical impact of differential item functioning analyses in a health-related quality of life instrument. Qual Life Res 18:1125–1130

37. Oort FJ, Visser MR, SprangersMA 2009 Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift. J Clin Epidemiol 62:1126–1137

38. Fava GA, BelaiseC2005A discussion on the role of clinimetrics and the misleading effects of psychometric theory. *J Clin Epidemiol* 58: 753–756
39. Wright JG, FeinsteinAR1992A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. *J Clin Epidemiol* 45:1201–1218
- Refer a

Figure legends

Figure 1 a. Fit to the Rasch model for the item “I have salt cravings”. The grey line depicts the expected scoring pattern as estimated by the model. The black dots are actual scoring from groups of patients with similar HRQoL levels (class intervals). The patients with the highest HRQoL level (far right) scores less than expected, the patients with the lowest HRQoL level (far left) scores better than expected; i.e. this item does not separate well between high and low HRQoL. **b.** DIF sex for the item “I have lost interest in sex”. The grey line depicts the expected scoring pattern estimated from the Rasch model. Men score better than expected, females worse; $p < 0.01$. This item showed similar results for DIF age. Patients below 50 years scored worse than expected, patients above 50 years better than expected; $p < 0.01$.

Figure 2. Person-item targeting for the revised 30-item AddiQoL. The upper half of the figure displays spread in AddiQoL scores (person location) for all patients. The lower half depicts item threshold distribution (item location). The item thresholds cover the range of HRQoL scores obtained by the patients, hereby minimizing the risk of floor- and ceiling effects.

Figure 3. AddiQoL-30 raw scores (range 30-120) in Norwegian patients (grey) and controls (black), in males (**a**) and females (**b**). Fatigue raw scores (range 8-32) in patients (grey) and controls (black) in males (**c**) and females (**d**). Error bars represent 95% CI.

Figure 4 a. Rasch-transformed AddiQoL-30 scores (range 0-100); and **b.** Rasch-transformed Fatigue scores (range 0-25); for each country; males gray, females white; median and interquartile range (boxes).

Table1. Patient characteristics

		Norway (n=107)	Italy* (n=157)	Germany (n=200)	Sweden (n=101)	Poland** (n=50)
Sex	Male (%)	39 (36.4)	54 (34.8)	53 (26.5)	36 (35.6)	10 (20)
	Female (%)	68 (63.6)	101 (65.2)	147 (73.5)	65 (64.4)	40 (80)
Age	18-29 (%)	6 (5.6)	16 (10.5)	15 (7.5)	6 (5.9)	NA
	30-39 (%)	17 (15.9)	42 (27.5)	35 (17.5)	18 (17.8)	NA
	40-49 (%)	29 (27.1)	43 (28.1)	58 (29.0)	22 (21.8)	NA
	50-59 (%)	32 (29.9)	27 (17.6)	42 (2.0)	27 (26.7)	NA
	60-69 (%)	22 (20.6)	17 (11.1)	29 (14.5)	17 (16.8)	NA
	> 70 (%)	1 (0.9)	8 (5.2)	21 (10.5)	11 (10.9)	NA
Comorbidity (autoimmune)	Present (%)	70 (65.4)	106 (68.4)	153 (76.5)	60 (59.4)	NA
	None (%)	37 (36.4)	46 (29.7)	47 (23.5)	41 (40.6)	NA
	Thyroid disease (%)	55 (50.9)	83 (52.5)	125 (62.5)	45 (44.6)	NA
	Type 1 diabetes (%)	13 (12.0)	6 (3.8)	20 (10)	9 (8.9)	NA
	Other***(%)	33 (30.6)	21 (13.4)	62 (31)	25 (24.8)	NA

* Italy: 16 patients were classified as APS2, these are included in comorbidities, but excluded from the thyroid and diabetes numbers. **Data regarding age and comorbidity are missing from Poland, and from a few Italian patients (sex, n=2); age (n=4) and comorbidities (n=2). ***Includes coeliac disease, hypoparathyroid disease, pernicious anemia, primary ovarian failure. NA; not available.

Table2. Overall fit to the Rasch model; the AddiQoL-30 results

	Norway (n=107)	Sweden (n=101)	Italy (n=156)	Germany (n=200)	Poland (n=50)	All countries (n=696)
Item Fit Residual (SD)	-0.55 (1.15)	-0.57 (1.45)	-0.51 (0.90)	0.01 (1.48)	-0.47 (0.9)	0.00 (2.7)
Person Fit Residual (SD)	-0.52 (1.06)	-0.62 (1.22)	-0.6 (1.08)	-0.48 (1.13)	-0.66 (1.2)	-0.51 (1.15)
χ^2 probability	0.81	0.66	0.88	0.73	0.92	0.73
Person Separation Index	0.84	0.87	0.83	0.91	0.85	0.86
Cronbach's α	ND	ND	ND	ND	ND	0.93

All countries also includes original UK data (n=82). Item Fit Residual: Mean item deviation from model estimations. Person Fit Residual: Mean person deviation from model estimations. A non-significant χ^2 probability implies that the hierarchical ordering of items and persons do not vary across the range of the scale. ND; not done.

Table3. Concurrent validity; correlation between AddiQoL30 scores and Fatigue scores with SF-36 and PGWB scores.

		Norway (n=107)		Sweden (n=101)	
		AddiQoL30	Fatigue	AddiQoL30	Fatigue
SF-36	Physical functioning	0.743	0.745	0.692	0.687
	Role Physical	0.689	0.729	0.717	0.661
	Bodily Pain	0.604	0.538	0.637	0.545
	General Health	0.802	0.775	0.768	0.692
	Vitality	0.753	0.748	0.837	0.803
	Social Functioning	0.685	0.676	0.603	0.585
	Role Emotional	0.433	0.418	0.466	0.411
	Mental Health	0.585	0.554	0.724	0.675
PGWB	Total score	0.816	0.797	0.785	0.706

Spearman's rho; All correlations were significant below the 0.01 level (two-tailed). SF-36; Short Form-36. PGWB; Psychological General Well-Being Index.

Supplementary Table1: AddiQoL items

Fatigue sub-dimension:

1. I feel good about my health
2. I can keep going during the day without feeling tired
3. Normal daily activities make me tired
4. I have to struggle to finish jobs
5. I have to push myself to do things
25. My ability to work is limited
28. I feel full of energy
29. I feel physically fit

Emotions sub-dimension:

12. I am relaxed
13. I feel low or depressed
14. I am irritable
15. I find it difficult to think clearly
26. I can concentrate well
27. I am happy
36. I cope well in emotional situations

Symptoms sub-dimension:

6. I lose track of what I want to say
10. I feel unwell first thing in the morning
16. I feel lightheaded
18. I sweat for no particular reason
19. I get headaches
20. I get nauseous
21. My joints and/or muscles ache
22. I have back pain
23. My legs feel weak
24. I worry about my health

Miscellaneous sub-dimension:

7. I sleep well
8. I feel rested when I wake up in the morning
11. I am satisfied with my sex life
34. I get ill more easily than others
35. I take a long time to recover from illnesses

Discarded items:

9. I need to get up during the night to pass water
17. I have salt cravings
30. Emotional stress makes me exhausted
31. I have lost interest in sex
32. I put on weight easily
33. I have dry skin